Actor-Critic を用いた知的照明システム

Intelligent Lighting Systems by Actor-Critic method

中村　康昭
Yasuaki NAKAMURA

Abstract: In this paper, we apply the Actor-Critic method which is another reinforcement learning method on an Intelligent Network Lighting System. Q-Learning is known to be effective if it is used on an environment with discreted action case. However, the action case of the environment we propose is continuous. In such environment, more suitable learning method is required.

1　はじめに

経験強化型の学習方法は様々なものが提案されており，現在の主流は TD 誤差学習である．TD 誤差学習の 1 つに，Q-Learning と Actor-Critic がある．Q-Learning はそれぞれの行動を評価するので，離散的な行動を必要とする際に有効であるとされる学習手法である．一方 Actor-Critic は行動の確率を評価するので，連続的な行動出力を必要とする際に有効である．本報告では，これらの学習法について説明し，Actor-Critic と Q-Learning を知的照明システムに適用させ，両者の違いを検証する．

2　学習手法

2.1　TD 誤差学習

TD 誤差学習では，現在の状況を観測し，その状況は自分が目標状態に到達するためにどの程度望ましい状態なのかを見積もる．見積もりと，実際に行動した際に得られた値との誤差を TD 誤差と呼ぶ．TD 誤差学習とは，TD 誤差を 0 に近づけていこうとする学習法である．TD 誤差は式 (1) で表される．

$$\begin{aligned}\sigma_t &= (\text{行動によって得られた値}) - (\text{見積もり}) \\ &= r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \qquad (1)\end{aligned}$$

r は，目標状態に達したときに与えられる報酬であり，$V(s_t)$ は行動前の状態の評価値，$V(s_{t+1})$ は行動後の状態の評価値である．$\gamma(0 \leqq \gamma \leqq 1)$ は割引率と呼ばれる係数である．これは未来の状態評価値にノイズや遅れがあることを考慮し，その値をそのまま信じずに割り引くためのものである．TD 誤差は状態 s_t において行動 a_t を出力した結果，どの状態へ遷移したかということから求まる．TD 誤差が正の時は現在の状態が見積もっていたよりも評価できるということであり，この場合，その時に選択された行動 a_t が選ばれる確率を大きくする．逆に，TD 誤差が負の時には見積もりよりも悪い状態だったということであり，行動 a_t が選ばれる確率を小さくする．評価値の更新は式 (2) に従う．

$$\begin{aligned}V(s_t) &\leftarrow V(s_t) + \alpha(TD - error) \\ &= V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)) \qquad (2)\end{aligned}$$

$\alpha(0 < \alpha \leqq 1)$ は学習率と呼ばる．これは現在の状態の評価値を次の状態の評価値にどれだけ近づけるかということを調節する．TD 誤差学習を用いた学習方法に Q-Learning や Actor-Critic がある．

2.2　Q-Learning

今，Fig. 1 のように状態遷移をする環境があるとする．

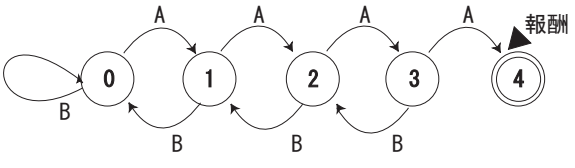

Fig. 1　例として用いるマルコフ決定過程

ここでは，状態が 5 つ定義されており，状態 3 において行動 A が選択されたときに状態 4，すなわち受理状態へ遷移している．また，初期状態は 0 から始まるものとし，受理状態まで到達すると再度 0 に移るものとする．それぞれの状態の初期評価値を 1 とし，学習率=0.5，割引率=1.0, 報酬=10 とする．最初の試行で，$0 \to 1 \to 2 \to 3 \to 2 \to 3 \to 4$ と状態遷移をし，次の試行では，$0 \to 1 \to 2 \to 3 \to 4$ と状態が遷移したとして，Q-Learning がどのような動作をするか説明する．

Q-Learning では，状態とその状態において選択することのできる行動とをセットとして考え，そのセットに対する評価を行う．この評価値を Q 値と呼ぶ．まず，Table 1 の左のような表を用意する．これは学習が行われる前の Q 値の表で，たとえば，2 行目を見ると，状態 0 の時の行動 A と行動 B に対する評価値がそれぞれ 0 となっている．状態 4 の評価値は 0 であり，報酬が 10

与えられる．2 回の学習によって，Q 値がどのように変動するかを示す．

Table 1　2 試行した際の Q 値の遷移

	A	B
0	0	0
1	0	0
2	0	0
3	0	0
4	0(10)	

→

	A	B
0	0	0
1	0	0
2	0	0
3	10	0
4	0(10)	

→

	A	B
0	0	0
1	0	0
2	5	0
3	10	0
4	0(10)	

最初の試行で，状態 3 において行動 A を出力すると報酬が加えられるので．この行動の評価があがる．次の試行では，状態 2 において行動 A を起こすと，状態 3 に遷移する．状態 3 では行動 A と行動 B で，評価値が違うが，Q-Learning では行動評価値の中でもっとも大きな値を取る．よって状態 2 の行動 A からみた状態 3 の評価値はそれまでの見積もりよりも高いので，状態 2 における行動 A の評価値が上がる．

2.3　Actor-Critic

続いて，Actor-Ciritc について，先ほどと同じ Fig. 1 の環境を用いて説明する．Actor-Critic では，状態評価と行動選択が独立した形で存在する．ここでは行動選択に正規乱数を用いるものとする．0〜4 のそれぞれの状態が，自分自身の評価値 (V) を持ち，また，それぞれの状態で，行動の確率を定める正規乱数の中心値 (μ) と標準偏差 (σ) を持つ．Table 2 に示すように，それらの初期値は各状態で同じ値を持つ．行動の選択では，各状態の μ と σ に基づく Fig. 2 のような頻度で発生する正規乱数を発生させ，それを元に行動を決定する．発生した乱数が正であれば行動 A を選択し，負の時には行動 B を選択する．そして行動後に状態を観測し，状態評価値を更新する．

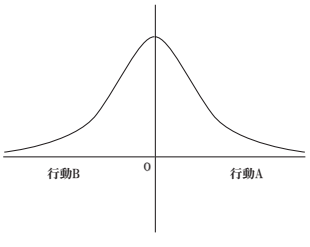

Fig. 2　発生する乱数の頻度と対応する行動

Table 2 に Actor-Critic を適用した際の V と μ の変動を示す．初期状態ではすべての状態で中心値 0，標準偏差 1 の正規乱数が発生する．行動を起こしたときに，行動後の評価値が高い場合，行動前の状態でその行動を選択することがよい結果へ動くことになり，行った行動の方向へ正規乱数の中心値を移動させる．

Table 2　Actor-Critic で用いる表の遷移

	V	μ	σ
0	0	0	1
1	0	0	1
2	0	0	1
3	0	0	1
4	0(10)		

→

	V	μ	σ
0	0	0	1
1	0	0	1
2	0	0	1
3	10	0.7	1
4	0(10)		

→

	V	μ	σ
0	0	0	1
1	0	0	1
2	5	0.5	1
3	10	0.7	1
4	0(10)		

最初の試行では受理状態へ遷移するときに TD 誤差が正となる．よって，状態 3 において，行動 A が選択されたとき，すなわち正の数が乱数として発生したときに，TD 誤差が正となる．このとき行動 A を導いた乱数の方へと中心値を移動させる．これによって状態 3 で正の乱数が発生する確率が高くなる．2 試行目には，状態 2 から状態 3 へ遷移したときにも TD 誤差が正となるので状態評価値が高くなり，中心値が正の方向へ更新される．

3　数値実験

3.1　Actor-Critic と Q-Learning の比較

3.1.1　対象問題と学習のパラメータ

Q-Learning と Actor-Critic の違いを検証するにあたり，知的照明システムを対象問題とした．このシステムは「人のいるところを X[lx] にせよ.」という共通の目的に従って動作する [1]．ライトは人のいる地点と，そこの明るさを検知できるものとする．

各学習法における行動選択手法は，Q-Learning では ε-greedy 選択を用いて，Actor-Critic では正規乱数に基づく行動選択を行った．

目標状態を 100[lx] として，誤差を ± 5[lx] とした．

状態数については，人のいる地点への最大照度を 300[lx] と考え，0〜300[lx] を誤差，すなわち今回は 5[lx] で分割して，s_0(0〜5)[lx]〜s_{60}(295〜300)[lx] とした．

パラメータは，Q-Learning と Actor-Critic でそれぞれ Table 3 のように設定した．

3.1.2　結果

Q-Learning と Actor-Critic を用いた場合の比較を Fig. 3 に示す．Q-Learning では，収束しても目標状態に至るまでに 30 ステップ程度は必要であったが，Actor-

Table 3　パラメータ

項目	Q-Learning	Actor-Critic
学習率 α	0.5	
割引率 γ	0.9	
選択確率 ϵ	0.2	/
中心値 μ	/	0
標準偏差 σ	/	1
状態評価値 s	0.1	
報酬 r	100	

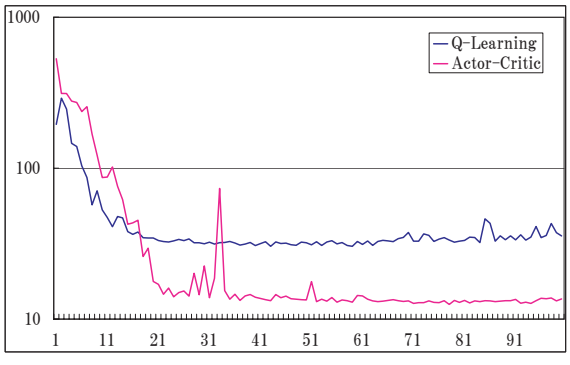

Fig. 3　Q-Learning と Actor-Critic の比較

Critic を用いると 10 ステップ程度で目標状態まで到達するようになった．

3.1.3　考察

知的照明システムでは，現状が目的の明るさに遠はライトの明るさを大きく変動さ，目的の明るさが近い時には，変動を小さくさせた方がよいと考えられる．一般に Q-Learning で連続的な行動出力を要求される環境を扱う場合，行動空間を分割して考える．しかし，変動の幅の大きなものから小さなものまで多くの行動を用意すると，行動の選択肢が増えるので，学習に時間がかかる．一方，行動の選択肢を少なくすると先ほどのようなきめ細かな制御は難しくなる．これに対して，Actor-Critic では，行動の選択と状態評価を分離しているため，行動の選択部分は確率を指定できるものであればよく，連続的な行動空間にも対応できる．

3.2　Actor-Critic のパラメータの検討

前節で，Actor-Critic の有効性が示せたので，パラメータを変化させ，それによって，挙動がどのように変化するかを調べた．

Actor-Critic で必ず必要となるパラメータは，学習率と割引率である．今回はこの学習率と割引率について実験を行った．この際，目標状態を人のいる地点を 100[lx] にすることとし，誤差を± 5[lx] とした．

$\alpha = 0.5$ とし，γ を変動させたときの結果を Fig. 4 に，$\gamma = 0.5$ とし，α を変動させたときの結果を Fig. 5 に示す．

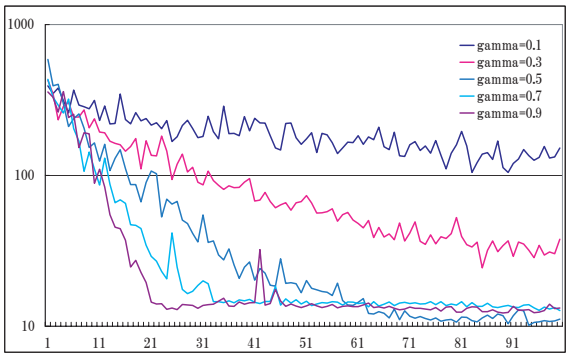

Fig. 4　学習率を変化させたときのステップ数の収束

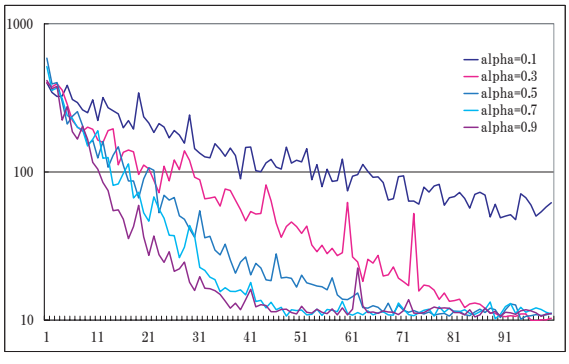

Fig. 5　割引率を変化させたときのステップ数の収束

γ が小さすぎると，収束が遅くなる．これはゴールから遠い状態における報酬の影響がでにくくなってしまっているためであると考えられる．α についても小さすぎると，報酬の影響が伝わりにくいため，収束が遅くなる．しかし，大きすぎると，収束は早いがその収束の値が大きくなってしまう．

4　まとめ

連続的な値を行動として選択したい状況であれば，Q-Learning よりも Actor-Critic を用いた方が効果的である．パラメータについては，割引率は 1 に近くても今回の対象問題では問題はないが，学習率については大きすぎると問題となることがわかった．今回の対象問題は，ライトを明るくするか，暗くするかのどちらかを選択するものであり，あとはその程度の問題となる．このような単純な問題では割引率はあまり考えなくてもよいようである．

参考文献

1) 富田浩二『知的ネットワークシステムの構築』　(2000 年度 修士論文)

2) 木村元, 宮崎和光, 小林重信『強化学習システムの設計指針』(計測と制御 Vol38,No.10,1/6, 1999)